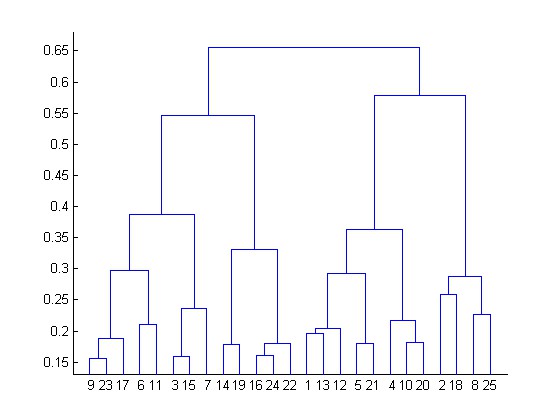
MACHINE LEARNING – WORKSHEET (CLUSTERING)

**Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.**

**1.** What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

a. 2 b. 4 c. 6 d. 8



**2.** In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers

2. Data points with different densities

3. Data points with round shapes

4. Data points with non-convex shapes

Options:

a. 1 and 2 b. 2 and 3 c. 2 and 4

d. 1, 2 and 4

e. 1, 2, 3 and 4

**3.** The most important part of

is selecting the variables on which clustering is based.

a. interpreting and profiling clusters b. selecting a clustering procedure

c. assessing the validity of clustering d. formulating the clustering problem

**4.** The most commonly used measure of similarity is the or its square. a. euclidean distance

b. city-block distance

c. Chebyshev’s distance

d. Manhattan distance

**5.** is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

a. Non-hierarchical clustering b. Divisive clustering

c. Agglomerative clustering d. K-means clustering

**6.** Which of the following is required by K-means clustering?

a. defined distance metric b. number of clusters

c. initial guess as to cluster centroids d. all answers are correct

**7.** The goal of clustering is to-

a. Divide the data points into groups

b. Classify the data point into different classes c. Predict the output values of input data points d. All of the above

**8.** Clustering is a-

a. Supervised learning

b. Unsupervised learning c. Reinforcement learning d. None

**9.** Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a. K- Means clustering

b. Hierarchical clustering c. Diverse clustering

d. All of the above

**10.** Which version of the clustering algorithm is most sensitive to outliers?

a. K-means clustering algorithm b. K-modes clustering algorithm

c. K-medians clustering algorithm d. None

**11.** Which of the following is a bad characteristic of a dataset for clustering analysis- a. Data points with outliers

b. Data points with different densities c. Data points with non-convex shapes d. All of the above

**12.** For clustering, we do not require- a. Labeled data

b. Unlabeled data c. Numerical data d. Categorical data

**Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly**.

**13.** How is cluster analysis calculated?

**Ans.** The k-Means method, which was developed by MacQueen (1967), is one of the most widely used non-hierarchical methods. It is a partitioning method, which is particularly suitable for large amounts of data.

* First, an initial partition with k clusters (given number of clusters) is created.
* Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated.
* If an object is detected whose distance to the center of gravity of the own cluster is greater than the distance to the center of gravity (centroid) of another cluster, this object is shifted to the other cluster.
* Finally, the centroids of the two changed clusters are calculated again, since the compositions have changed here.
* These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (center of the cluster) (optimal solution)

**14.** How is cluster quality measured?

**Ans.** Usually clustering algorithms are non-supervised which often means that the correct  partition into classes is not know and the number of classes is also unknown. In these cases there are measures of cluster validity based in maximizing inter-cluster distances and minimizing intra-cluster distances. There is a vast scientific literature on cluster validity. If in your case you know the correct partition then you can use a distance measure between partitions to measure how similar to the correct clustering is the clustering obtained by the algorithm. The choice of metric rather depends on what you consider the purpose of clustering to be. Personally I think clustering ought to be about identifying different groups of observations that were each generated by a different data generating process. So I would test the quality of a clustering by generating data from known data generating processes and then calculate how often patterns are misclassified by the clustering. Of course this involved making assumtions about the distribution of patterns from each generating process, but you can use datasets designed for supervised classification. Here you have a couple of measures, but there are many more:

SSE: sum of the square error from the items of each cluster.

Inter cluster distance: sum of the square distance between each cluster centroid.

Intra cluster distance for each cluster: sum of the square distance from the items of each cluster to its centroid.

Maximum Radius: largest distance from an instance to its cluster centroid.

Average Radius: sum of the largest distance from an instance to its cluster centroid divided by the number of clusters.

**15.** What is cluster analysis and its types?

**Ans.** Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets.

Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal.

In other words, the clusters are regions where the density of similar data points is high. It is generally used for the analysis of the data set, to find insightful data among huge data sets and draw inferences from it.

The types of Clustering Methods

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.

* Density-Based Clustering:

In this method, the clusters are created based upon the density of the data points which are represented in the data space. The regions that become dense due to the huge number of data points residing in that region are considered as clusters.

The data points in the sparse region (the region where the data points are very less) are considered as noise or outliers. The clusters created in these methods can be of arbitrary shape.

* Hierarchical Clustering

Hierarchical Clustering groups (Agglomerative or also called as Bottom-Up Approach) or divides (Divisive or also called as Top-Down Approach) the clusters based on the distance metrics. In Agglomerative clustering, each data point acts as a cluster initially, and then it groups the clusters one by one.

Divisive is the opposite of Agglomerative, it starts off with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between the clusters depending on the criteria of the linkage. The clustering of the data points is represented by using a dendrogram.

* Fuzzy Clustering

In fuzzy clustering, the assignment of the data points in any of the clusters is not decisive. Here, one data point can belong to more than one cluster. It provides the outcome as the probability of the data point belonging to each of the clusters. One of the algorithms used in fuzzy clustering is Fuzzy c-means clustering.

* Partitioning Clustering

This method is one of the most popular choices for analysts to create clusters. In partitioning clustering, the clusters are partitioned based upon the characteristics of the data points. We need to specify the number of clusters to be created for this clustering method. These clustering algorithms follow an iterative process to reassign the data points between clusters based upon the distance.

* Grid-Based Clustering

In grid-based clustering, the data set is represented into a grid structure which comprises of grids. The overall approach in the algorithms of this method differs from the rest of the algorithms. They are more concerned with the value space surrounding the data points rather than the data points themselves. One of the greatest advantages of these algorithms is its reduction in computational complexity. This makes it appropriate for dealing with humongous data sets. After partitioning the data sets into cells, it computes the density of the cells which helps in identifying the clusters.